УДК 004.822

СПОСОБ АВТОМАТИЧЕСКОГО ОПРЕДЕЛЕНИЯ ТЕМАТИКИ НАУЧНОГО ТЕКСТА

Аюшеева Н.Н., Гомбожапова Т.Н., Доржаев Т.В.

Восточно-Сибирский государственный университет технологий и управления, Улан-Удэ, e-mail: donir@rambler.ru

Данная статья посвящена проблеме автоматического определения темы и тематики научного текста для построения его семантической сети. Предложенный в работе способ основан на анализе частотных характеристик терминов и отношений в тексте, а также на рассмотрении композиционной структуры научного документа. Под частотными характеристиками понимаются частота встречаемости и ее ранг. Выделение композиционной структуры текста основано на представлении научного текста его содержательно-смысловыми блоками, например, блок постановки и понимания проблемы, блок изложения варианта решения проблемы и т.п. Определение тематики текста можно разбить на четыре последовательных этапа: выбор ключевых терминов анализируемого текста; группирование выбранных ключевых терминов; выделение отношений с наибольшим весом; формирование тематики текста. Данный способ будет использован для расчета весовых коэффициентов терминов и отношений при построении семантической сети научного текста.

Ключевые слова: тема текста, тематика текста, научный текст, семантическая сеть, весовой коэффициент вершины

AN AUTOMATIC SCIENTIFIC TEXT TOPIC IDENTIFICATION METHOD

Ajusheeva N.N., Gombozhapova T.N., Dorzhaev T.V.

East Siberia State University of Technology and Management, Ulan-Udje, e-mail: donir@rambler.ru

This paper is referred to the task of automatic scientific text topic identification for construction a semantic network. The method is based on analysis of term and relationship frequency characteristics, and this method is studied a scientific text compositional structure. Here, a frequency characteristics are the term or relationship frequency and their rank. The scientific text compositional structure consist of the text semantic content-blocks, such as Problem block, Solution block, etc. The proposed method contains four sequential steps: select the keywords; group of selected keywords; distinguish the most weighed relationships; form the topics. In future this method will used to calculate term and relationship weights, when the scientific text semantic network will has been built.

Keywords: text topic, text subject, scientific text, semantic network, weight of term

Выделение темы текста представляет собой неформализуемый процесс, который выполняется экспертом на основе его знаний и зависит от его интуиции и опыта. В связи с этим определение текста является слабоструктурированной задачей. Несмотря на это, решению задачи автоматического определения темы и тематики текстового документа посвящено большое количество исследований и разработок. Однако результаты выполненных изысканий чаще всего носят локализованный характер, т.е. ограничены конкретным естественным языком, предметной областью, сферой применения. Другая причина продолжения исследований в избранном направлении заключается в возможности использования дополнительного лингвистического обеспечения для улучшения качественных характеристик определения тематики, например, таких как рубрикаторы, тезаурусы, онтологии.

В работе [1] представлены методы построения семантической сети научного текста. Важной задачей, решаемой при построении семантической сети текстового документа, является задача определения значимости терминов текста, которые влияют на определение его смысла. К основным критериям значимости отнесены [2]:

– частота встречаемости термина в документе;

– категория текста, в которую входит термин;

– содержательно-смысловой блок, в котором термин встречается.

Категория текста как смысловая часть текста отражает один из компонентов коммуникативного акта, в число которых входит предмет речи; субъект(-ы) речи, то есть автор(-ы) текста в целом; оценочная точка зрения субъекта; его эмоционально-психологический настрой; пространство и время как неотъемлемые атрибуты ситуации, в которой порождается текст; адресат общения. Соответственно выделяются текстовые категории темы, субъекта (авторизации), оценочности, тональности (текстовой модальности), текстового пространства и времени, адресата. Для построения адекватной семантической сети важно обратить внимание на текстовую категорию темы. В работе [2] вклад в значение весового коэффициента

термина принят равным 1, если термин отражает тему текста, и 0 – в противном случае. Таким образом, необходимо определить тематику научного текста для того, чтобы повысить вес значимых терминов текста.

Основные понятия и определения

Рассмотрим понятие «тема текста». Тема – существенный и необходимый признак всякого текста. Тема текста – это предмет обсуждения в тексте, номинативно выраженное содержательное ядро целого текста, сопоставимое с авторским замыслом в целом [3]. Тема сохраняет свое единство на протяжении текста, обеспечивая его целостность. В совокупности со своим смысловым предикатом тема текста образует тезис целого текста – выражение его основной мысли. В крупном тексте тема делится на подтемы и субподтемы, отражающие содержание относительно самостоятельных частей целого текста. В работе [4] рассматривается, что текст состоит из компонентов высказывания, таких как тема и рема. Тема – это известная исходная информация, от которой развивается рема, то есть новая информация. Новая информация содержит основное сообщение, выделяется логическим ударением и обычно находится в конце предложения. При рассмотрении тема-рематического строения текста исходят из коммуникативной структуры предложения, которая имеет план выражения и план содержания. Предложение – сочетание двух смысловых центров, находящихся в тесном единстве: первый смысловой центр обозначает то, о чем говорится в высказывании – тему (предмет речи), второй смысловой центр обозначает то, что об этом говорится – рему (содержание речи).

Тема определяется по следующим параметрам:

– тема – предмет речи;

– тема выделяется логическим ударением;

– тема определяется порядком слов в предложении, как правило, в начале.

Следует заметить, что нередко возникают трудности при определении тема-рематической организации текста, так как иногда невозможно сказать однозначно, что составляет тематический или рематический элемент сверхфразового единства, а порой размыты и сами границы сверхфразового единства. Тема-рематическая организация любых текстов зависит от жанра, стиля и композиционных особенностей текста. Научная коммуникация, как известно, характеризуется своими особенностями, а именно стремлением к точности и логичности мысли. В данном стиле явно преобладает функция сообщения, фиксации результатов познания мира. Знание представлено в строго аргументированной форме, особое место уделяется ходу логических рассуждений. Особенности стиля отражаются и на тема-рематической организации научных текстов.

Из вышесказанного следуют два заключения. Во-первых, тематика текста должна быть представлена некоторой совокупностью тем и/или подтем. Во-вторых, при определении тематики научного текста нужно учитывать особенности этого стиля текстового изложения.

Обзор известных методов и способов определения тематики текста

Исследования в области определения тематики текстовой информации проводятся со времени глобального распространения электронной информации. Необходимость автоматизации данного процесса становится острее по мере роста объемов информации, которую следует обработать – прочитать, проанализировать, изучить, найти. Согласно [5] частными задачами определения тематики текста являются задачи выделения терминов (Term Assingmemt), извлечения ключевых фраз (Keyphrase Extraction), теггирования документов (Document Tagging). В своем исследовании автор предлагает подход автоматического определения тематики документа на основе использования онтологических знаний. При этом используется собственная онтология, построенная по статьям Википедии. Предложенный подход ориентирован на определение соответствия некоторого текстового документа одной тематике из заданного множества тематик. В работе [6] также используется онтология для определения тематик веб-страниц. Здесь выделяются ключевые слова из заранее определенного множества тэгов, которые затем ищутся среди концептов онтологии. Далее полученные знания используются для определения сходства между ключевыми словами. Другой класс известных методов определения темы текста образуют методы, основанные на кластеризации текстовой информации [7]. Эти методы основаны на выделении потоков терминов и распределении их по кластерам. Тематика текстовой информации определяется путем анализа выделенных кластеров. Среди отечественных разработок внимание заслуживают системы автоматического определения тематики текста «Семантическое зеркало», эксперт тематики «ExTheme», система семантического анализа текста «Advego». Система «Семантическое зеркало» определяет тему текста по созданному лингвистами рубрикатору, который содержит около 3000 рубрик,

каждая из которых может включать от сотни до несколько тысяч терминов. В настоящее время в модуле насчитывается более 600000 терминов. Система при определении, к какой рубрике принадлежит текст, опирается на следующие факторы: вес термина; длина термина; количество вхождений термина; общее количество терминов в тексте; местоположение термина; особенности употребления термина и другие. Система экспертной оценки «ExTheme» так же, как и «Семантическое зеркало» прибегает к помощи рубрикатора. В качестве рубрикатора система использует Яндекс-каталог. Система семантического анализа текста «Advego» в качестве тематики текста предлагает использовать семантическое ядро текста, т.е. фразы или слова с высокой частотой встречаемости. Резюмируя сказанное, следует отметить, что применение вспомогательного лингвистического обеспечения, каковыми являются онтологии и рубрикаторы, значительно ограничивает применимость метода вследствие сложности разработки онтологий и рубрикаторов. Однако точность таких методов значительно выше.

Предлагаемый способ определения тематики научного текста

Для автоматического определения тематики текста необходимо выполнить выделение терминов текста, определить их статистические характеристики и выявить отношения между терминами. Данные процедуры подробно описаны в работах [1, 2].

Определение тематики текста схематично можно разбить на четыре последовательных этапа:

– выбор ключевых терминов и отношений анализируемого текста;

– группирование выбранных ключевых терминов;

– выделение отношений с наибольшим весом;

– формирование тематики текста.

Первый этап заключается в формировании множеств ключевых терминов T^K. Указанные множества образуются на основе анализа весовых коэффициентов терминов текста:

$$T^K = \left\{ t_i \middle| w_i^t \geq \mu, \quad i \in \{1...N\} \right\}, \qquad (1)$$

где t_i – термин текста; w_i^t – весовой коэффициент термина, μ – пороговое значение весовых коэффициентов, N – количество терминов текста.

Пороговое значение весовых коэффициентов вычисляется по формуле

$$\mu = \sqrt{\frac{1}{N}\sum_{i=1}^{N} w_i^t}. \qquad (2)$$

Выбор данной формулы обоснован тем, что в этом случае пороговый коэффициент обеспечивает достаточно постоянное значение доли выделенных понятий и терминов в общем числе понятий и терминов текста и наименьшее значение углового коэффициента зависимости количества ключевых слов от объема текста.

На втором этапе ключевые термины группируются по видам отношений. В данной работе остановимся на трех видах отношений: отношение синонимии, отношение «часть-целое», отношение «род-вид». Данные отношения являются универсальными. Они присущи не только научной терминологии, но и элементам любого естественного языка [8]. В зависимости от деления тезаурусных функций на парадигматические и синтагматические отношения синонимии и «род-вид» являются парадигматическими. Отношение «часть-целое» не является ни парадигматическим, ни синтагматическим. Его позиционируют как «ситуативно открытое». Согласно классификации отношений на языковые, понятийные и предметные отношение синонимии относят к языковым, родовидовые отношения и «часть-целое» – к понятийным. С точки зрения иерархичности, синонимия является не иерархичным отношением, а «род-вид» и «часть-целое» – иерархичные отношения. Выбор для рассмотрения указанных трех видов отношений является вполне достаточным, а множество выбранных отношений репрезентативным.

Группирование терминов по видам отношений выполняется на основе онтологических знаний. В результате формируется три группы терминов *SIN*, *AKO* (*A kind of*), *HasPart*.

Группа *SIN* представляет собой множество подмножеств ключевых терминов sin_i:

$$SIN = \{\sin_i | i = 1...n,\ n – \text{количество различных синонимов}\}. \qquad (3)$$

Два ключевых термина принадлежат одному подмножеству *sin*, если они являются синонимами:

$$sin_i = \{x_j | x_j \text{ синоним } x_{j+1}, j = 1...k_i - 1,\ k_i – \text{количество синонимов}\}. \qquad (4)$$

В каждом подмножестве синонимов выбирается один термин, у которого вес является наибольшим. Если терминов с наибольшим весом более одного, то выбираются все.

Группа *AKO* – это множество пар ключевых терминов (t_i, t_j), в которых t_i, t_j являются терминами, находящимися в отношении «род-вид»:

$$AKO = \{(t_i, t_j)|i, j = 1...m, m \text{ – количество пар терминов}\}. \quad (5)$$

Группа *HasPart* – это множество пар терминов (t_i, t_j), в которых t_i, t_j являются терминами, находящимися в отношении «часть-целое» (6). Причем термин t_i – часть, должен принадлежать множеству ключевых терминов. А термин t_j – целое, определяется по онтологии предметной области.

$$HasPart = \{(t_i, t_j)|i, j = 1...m, m \text{ – количество пар терминов}\}. \quad (6)$$

На третьем этапе выполняется отбор отношений R^K, в которых участвуют ключевые термины. Из отобранных отношений оставляются отношения, весовой коэффициент которых больше вычисленного по формуле (7) порогового значения веса отношений.

На последнем этапе определяется тематика текста по следующим правилам:

– в отношении r_i выполняется замена термина, имеющего синоним с большим значением веса; если вес термина больше веса своих синонимов, то он остается на своем месте;

– в отношении r_i термин, который является видом родового термина с большим весом, заменяется на термин, являющийся родом для него;

– если во множестве ключевых терминов имеется несколько терминов, являющихся частью одного и того же целого, то в отношении, в котором встречается термин-часть, выполняется замена этого термина на словосочетание «компоненты термин-целое».

Практическое применение способа определения тематики научного текста

Программный модуль определения тематики текста должен стать структурным компонентом модуля вычисления весовых коэффициентов семантической сети. На рис. 1 представлена обновленная структура модуля вычисления весовых коэффициентов семантической сети.

На вход модуля вычисления весовых коэффициентов семантической сети поступает множество терминов *T*, множество отношений *R* и непосредственно сам текст *D*. На выходе получаем взвешенную семантическую сеть *S*.

Модуль вычисления весовых коэффициентов вершин и дуг семантической сети включает компонент вычисления веса вершин, компонент вычисления веса дуг, нечеткий регулятор определения содержательно-смыслового блока термина или отношения и компонент определения тематики. Нечеткий регулятор определения содержательно-смыслового блока термина или отношения получает от компонента вычисления веса вершин термин *t*, а от компонента вычисления веса отношений – отношение *r*. Нечеткий регулятор обратно возвращает название содержательно-смыслового блока B, в котором находится анализируемый термин/отношение. Компонент определения тематики получает множество взвешенных терминов текста *T* от компонента вычисления веса терминов и множество взвешенных отношений текста *R* от компонента вычисления

Рис. 1. Обновленная структура модуля вычисления весовых коэффициентов семантической сети

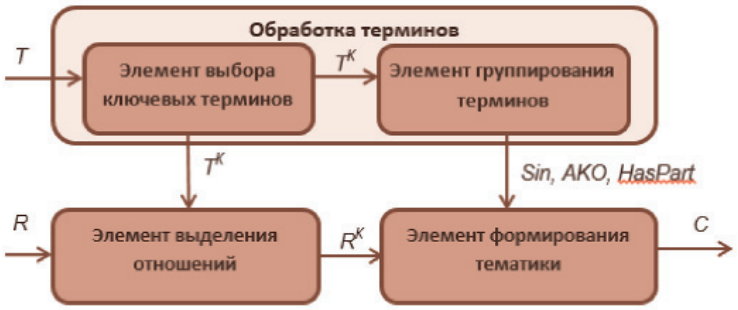

Рис. 2. Структура компонента определения тематики

веса отношений. Обратно он возвращает тематику текста *C*, которая позволяет корректировать вес терминов и отношений.

Рассмотрим структуру компонента определения тематики текста (рис. 2).

В соответствии с определенными выше этапами определения тематики выделен элементный состав компонента. Связи между элементами являются информационными и представляют собой потоки данных. Результатом работы элемента выбора ключевых терминов является множество ключевых терминов T^K. В результате работы элемента группирования терминов формируются множества синонимов *Sin*, терминов – участников отношения «род-вид» *AKO*, пар терминов отношений «часть-целое» *HasPart*. Элемент выделения отношений формирует множество отношений R^K, а элемент формирования тематики выдает конечный результат.

Заключение

Предложенный в данной статье способ автоматического определения тематики текста ориентирован на его использование для построения семантической сети научного текста, так как весовой коэффициент терминов и отношений, учитываемый при определении тематики, рассчитан на основе анализа композиционной структуры научного текста. Применение онтологических знаний здесь оправдано тем, что сам метод построения семантической сети текста также базируется на анализе концептуальных связей онтологий предметных областей. Для выполнения вычислительных экспериментов разработано программное приложение. Реализованный в программе предложенный способ показывает вполне приемлемые результаты. Сравнение с результатами, полученными существующими системами, демонстрирует, что предложенный метод выдает более качественные результаты.

Список литературы

1. Аюшеева Н.Н., Гомбожапова Т.Н. Разработка методов построения семантической сети текста: моногр. – Улан-Удэ: Изд-во ВСГУТУ, 2016. – 124 с.

2. Аюшеева Н.Н., Кушеева Т.Н. Способ вычисления весовых коэффициентов вершин семантической сети научного текста // Фундаментальные исследования. – 2012. – № 6–3. – С. 626–631.

3. Стилистический энциклопедический словарь русского языка / под ред. М.Н. Кожиной. – 2-е изд. – М.: Флинта: Наука, 2006. – 696 с.

4. Даркулова К.Н., Мамет А. Трудности определения тема-рематической связи в научном тексте [Электронный ресурс]. – Режим доступа: http://www.scienceforum.ru/2014/476/71 (дата обращения 04.07.2016).

5. Hassan M. Automatic Document Topic Identification Using Hierarchical Ontology Extracted from Human Background Knowledge. – Canada, 2013. – 129 p.

6. Rathore F.S., Roy D. Ontology based Web Page Topic Identification // International Journal of Computer Applications. – January 2014. – № 85(6). – P. 35–40.

7. Hossein Shahsavand Baghdadi, Bali Ranaivo-Malançon. An Automatic Topic Identification Algorithm // Journal of Computer Science. – 2011. – № 7 (9). – P. 1363–1367.

8. Никитина С.Е. Семантический анализ языка науки. – М.: Наука, 1987. – 144 с.

References

1. Ajusheeva N.N., Gombozhapova T.N. *Razrabotka metodov postroenija semanticheskoj seti teksta* [The creating of text semantic network building methods]. Ulan-Udje. ESSUTM. 2016. 124 p.

2. Ajusheeva N.N., Kusheeva T.N. *Fundamentalnye issledovanija,* 2012, no.6 (3), pp. 626–631.

3. *Stilisticheskij jenciklopedicheskij slovar russkogo jazyka* [Stylistic Encyclopedic Dictionary of the Russian language]. Moscow. 2006. 696 p.

4. Darkulova K.N., Mamet A. Available at: http://www.scienceforum.ru/2014/476/71 (accessed 4 July 2016).

5. Hassan M. Automatic Document Topic Identification Using Hierarchical Ontology Extracted from Human Background Knowledge. Canada, 2013. 129 p.

6. Rathore F.S., Roy D. Ontology based Web Page Topic Identification. International Journal of Computer Applications 85(6): 35–40, January 2014.

7. Hossein Shahsavand Baghdadi, Bali Ranaivo-Malançon. An Automatic Topic Identification Algorithm. Journal of Computer Science 7 (9): 1363–1367, 2011.

8. Nikitina S.E. *Semanticheskij analiz jazyka nauki* [Semantic Analysis of the Science Language]. Moscow. 1987. 144 p.